3Н

УДК 681.3

А.О. Никоненко
Київський національний університет ім. Т. Шевченка, м. Київ, Україна
andrey.nikonenko@gmail.com

Огляд комп’ютерно-лінгвістичних методів обробки природномовних текстів

У статті проведено дослідження методів автоматичного аналізу природномовних текстів. Детально розглянуто чотири основних підходи: морфологічний, статистичний, синтаксичний, семантичний. Описані можливості їх використання у розв’язанні прикладних задач та специфіка застосування. Як приклади наведено перелік прикладних систем, що використовують перераховані методи.

Вступ

Задача даного дослідження – провести огляд сучасних комп’ютерно-лінгвістичних методів аналізу текстової інформації з метою визначення найбільш перспективних напрямків проведення подальших робіт у цій області. На сьогодні існує необхідність у створенні чіткої класифікації лінгвістичних методів та розбиттю існуючих підходів на класи еквівалентності для створення єдиної термінології типових лінгвістичних шаблонів. Створення шкали таких методів дозолить чітко відрізняти одні підходи від інших, буде сприяти підвищенню чіткості та однозначності у використанні термінів, дозволить вивести набір ознак, за якими можна буде чітко розмежувати одні методи від інших. Фіксація лінгвістичних методів, їх особливостей, способу реалізації та типових застосувань дозволить надати всій комп’ютерній лінгвістиці більш структурованого характеру. На основі запропонованої класифікації можливе подальше створення класів мішаних методів для вирішення задач, що знаходяться на межі двох (або більше) типів.

Морфологічний підхід

У даному пункті нами буде наведено короткий огляд множини сучасних методів, що спираються на певні морфологічні перетворення тексту: нормування вхідного тексту, розширення запиту списком словоформ, аналіз граматичних класів слів та інше.

Повнотекстовий пошук в документах. Одним з основних завдань, що виникають при роботі з повнотекстовими базами даних, є завдання пошуку документів за змістом. Однак традиційні засоби контекстного пошуку, що орієнтуються на входження слів у документ, часто не забезпечують коректного вибору інформації за запитом користувача. Основна проблема полягає в складності точного формулювання запиту – підбору ключових слів, які належить шукати в тілі документу. Це може бути пов’язано з рядом причин, серед них: недостатнє знання користувачем термінології предметної області, складнощі з визначенням меж своїх інтересів, наявністю в мові багатозначних і синонімічних слів, а також орфографічними помилками в написанні слів, які можуть зустрічатися як у текстах, так і в самому запиті.

Технологія пошуку з орфографічними помилками, запропонована компанією RCO [1], дозволяє розширювати запит близькими за написанням словами, що містяться в колекції документів, за якими ведеться пошук. Такий метод може бути застосовано,

якщо документи, за якими ведеться пошук, містять слова з орфографічними помилками, або при наявності сумніву в правильності написання слів у запиті – імен, назв і т.д. Наприклад, запит Інкомбанк може бути розширений словами: інкомбан, інкобанки, вінкомбанку. А якщо користувач забув точну назву медичного препарату Іпроніазід, то можна задати що-небудь схоже, наприклад, імпронізід – потрібні документи буде знайдено.

Алгоритм, використаний для реалізації пошуку схожих слів, заснований на системі асоціативного доступу до слів, що містяться в текстовому індексі повнотекстового сховища документів. Швидкість пошуку пропорційна логарифму від числа індексованих слів і становить менше однієї секунди при індексі в кілька мільйонів слів (такий повнотекстовий індекс відповідає кільком гігабайтам повнотекстових документів). Пошук здатний знайти всі лексикографічно близькі слова, що відрізняються замінами, пропусками і вставками допустимої кількості символів.

Лексико-граматичний аналіз. Завдання лексико-граматичного аналізу – автоматично розпізнати, до якої частини мови належить кожне слово в тексті. В [2] автори пропонують наступний метод розбору речення з проставлянням лексико-граматичних класів кожному слову. Дано речення:

When you access the BIB record you want, you can print the screen, write down any information you need, or select the item if you are placing a hold.

Після проставляння лексико-граматичних класів воно стане виглядати так (WRB, PPSS, VB та інші вказують на класи слів):

When/WRB you/PPSS access/VB the/AT BIB/NN record/NN you/PPSS want/VB ,/, you/PPSS can/MD print/VB the/AT screen/NN ,/, write/VB down/RP any/DTI information/NN you/PPSS need/VB ,/, or/CC select/VB the/AT item/NN if/CS you/PPSS are/BER placing/VBG a/AT hold/NN ./.

Відповідно до викладеного в роботі [3] матеріалу, існує два види алгоритмів для розв'язання даної задачі: ймовірнісно-статистичні та засновані на продукційних правилах, що оперують словами та їх кодами.

Ймовірнісно-статистичні потребують джерело інформації для навчання, відповідно до роботи [4], зазвичай використовується два типи джерел:

1. Словник словоформ мови, що містить інформацію про відповідність словоформ до множини лексико-граматичних класів. Для кожного лексико-граматичного класу словоформи вказується частота його використання щодо інших лексико-граматичних класів даної словоформи. Частота зазвичай підраховується на розміченому вручну корпусі текстів.

2. Інформація про частоту використання всіх можливих послідовностей лексико-граматичних класів. Залежно від того, як представлена дана інформація, розділяють біграмну, триграмну і квадріграмну модель.

Дана інформація опрацьовується програмою, що використовує статистичні алгоритми, найчастіше алгоритм прихованих ланцюгів Маркова [4] для знаходження найбільш ймовірного лексико-граматичного класу для кожного слова в реченні.

Алгоритми, засновані на продукційних правилах, використовують правила, зібрані автоматично з корпусу текстів [5] або підготовлені кваліфікованими лінгвістами [6].

Використання обох підходів дає приблизно однаковий результат [7], [8].

Нечіткий пошук в інтернеті. Корисний результат, що досягається при використанні описаного в [1] способу пошуку, полягає в підвищенні точності пошуку при збереженні його високої повноти, а також у зниженні навантаження на пошукову машину.

Даний спосіб засновано на використанні лінгвістичних знань про граматику тієї природної мови, на якій формулюється пошуковий запит, а саме синтаксичних зв'язків між словами пошукового запиту для вибору оптимального відображення на мову за-

питів пошукової машини. А також, при відсутності результатів пошуку документів з використанням виразу, для формування послідовності пошукових виразів з меншим ступенем строгості пошукових обмежень і з максимально можливим збереженням сенсу початкового запиту, для забезпечення послідовного підвищення повноти пошуку з мінімальною втратою точності. Відповідність операторів мови запитів синтаксичним зв'язкам між словами встановлюється на підставі того принципу, що більш сильно пов'язані в запиті слова повинні шукатися на ближчій відстані в тексті і з більш жорсткими обмеженнями на допустимі граматичні форми.

Статистичний підхід

У даному пункті ми розглянемо методи, що отримали назву асоціативно-статистичні [10] на противагу чистим математично-статистичним методам. Дані методи використовують лінгвістичні моделі для опрацювання результатів статистичного аналізу тексту, що дозволяє отримати дещо кращі результати, порівняно з чистою статистичною обробкою текстових документів.

Тематичний аналіз. В основі методу лежить уявлення змісту тексту у формі асоціативно-семантичної мережі, вузли якої представлені множиною понять, що часто зустрічаються в тексті – слова та стійкі словосполучення, з числа яких виключаються загальновживані слова. Вузли мережі асоціативно пов'язані між собою з різною силою, причому сила зв'язку корелює з частотою спільного входження понять в речення тексту. Семантична мережа може бути автоматично побудована на базі корпусу текстів і використана згодом як модель предметної області для аналізу невідомих документів.

У моделі процесу породження тексту [9] поява речення вважається обумовленою активацією одного вузла мережі, що знаходиться у фокусі уваги і представляє тему висловлювання. Поява інших слів у реченні обумовлена їх зв'язками з темою, що задіяні в мережі на момент породження тексту. Враховуючи надфразову зв'язність повідомлення в цілому, вважається, що найбільш ймовірно зумовлення теми висловлювання темою або ремою попереднього, що відображає збереження фокусу уваги або його перемикання на пов'язаний вузол мережі [10]. В результаті, породження тексту можна представити як марковський процес, стани якого відповідають реченням, а вірогідність переходів між ними обумовлюється силою зв'язків елементів семантичної мережі.

Якщо існує декілька еталонних мереж, які представляють тематичні класи близьких за змістом документів, то можна класифікувати новий текст, визначивши ймовірність його породження кожною мережею.

Реферування

1. Побудова рефератів базується на методі, близькому до описаного вище. В основі методу лежить побудова асоціативно-семантичної мережі, що представляє собою орієнтований граф, вершинами якого служать значимі теми, виділені в аналізованому тексті, а дугами – зв'язки між ними. З кожною вершиною зв'язана вага (значимість) і частота згадування теми, а з кожною дугою – вага (сила) зв'язку та частота підкріплення зв'язку в тексті.

Крім частоти згадування в тексті, кожній темі встановлюється вага від 1 до 100, що відображує її значимість відносно інших тем. Користувач може задати мінімальний поріг по вазі, нижче якого теми не включаються у семантичну мережу.

Асоціативні зв'язки між темами будуються на основі частоти їхньої спільної появи в одному реченні. Користувач може задати мінімальний поріг частоти, нижче якого зв'язки відкидаються. У кінцевому представленні зв'язок перетворюється у дві протилежно спрямовані дуги графа, яким присвоюються ваги від 1 до 100, що відображають умовну ймовірність згадування першої теми спільно з другою – силу зв'язку.

Після побудови мережі для кожної теми видається реферат, що представляє найбільш інформативні фрагменти тексту, у яких дана тема згадувалася. Загальний реферат тексту представляє компіляцію найбільш інформативних фрагментів по ключових темах [11].

2. Інший підхід до побудови рефератів запропоновано в системі Inxight Summarizer [12]. Дана комерційна система була створена в Дослідницькому центрі Xerox в Упало Альто. В основі її алгоритмів реферування лежить принцип виділення найбільш статистично вагомих речень тексту та використання слів-підказок. Також розробники системи створили один з найбільш точних алгоритмів оцінки якості реферату. Паралельне використання відразу декількох широко відомих статистичних алгоритмів реферування та безпосередній зв'язок між результатами їх роботи й алгоритмом оцінки якості реферату забезпечили тривалий успіх системи.

Авторубрикація та класифікація. Теоретичне доведення даного методу дано в роботі [13]. Як і більшість описаних в цьому пункті методів, він функціонує на базі асоціативно-семантичної мережі, що будується статистичними методами. Метод базується на припущенні, що безліч текстів, які відносяться до одного класу, породжуються на основі однієї семантичної мережі – еталона. Тоді завдання віднесення невідомого тексту до відповідного класу зводиться до визначення ймовірностей породження тексту на основі кожної з еталонних мереж. Для використання даного методу потрібно попередньо провести налаштування параметрів мереж-еталонів на базі навчальної вибірки текстів, розміченої по класах.

Основу методу складає спосіб виділення понять мережі (тобто слів та словосполучень), для цього використовується статистичний алгоритм, заснований на аналізі частоти входження в текст ланцюжків слів різної довжини та їх взаємного входження одне в одного, описаний в [13].

Загалом оцінка ваг зв'язків розраховується на базі відношення частоти спільного входження понять у речення тексту, нормованого за кількістю понять в кожному з речень, до частоти входження поняття в текст (крім повторів в одному реченні).

Синтаксичний підхід

Огляд технологій, представлених в даному пункті, в основному базується на інформації, отриманій з роботи [14] та деяких інших робіт тих же авторів. Основна проблема, що існує при застосуванні синтаксичного підходу – дуже складна структура речень у флективних мовах (російська, українська, польська…) на відміну від достатньо простої структури в аналітичних мовах (англійська, французька, італійська, болгарська...). Тому задача повного синтаксичного аналізу для флективних мов досі не розв'язана. Далі ми розглянемо підходи до комп'ютерно-лінгвістичної обробки текстів, що базуються на неповному синтаксичному розборі структури речень.

Розв'язання омонімії. Метою синтаксичного розбору є побудова дерева синтаксичних залежностей між словами речення. У випадку вдалого розбору все речення буде звернуто у повнозв'язне дерево з єдиним коренем.

Оскільки одна словоформа може відповідати декільком граматичним формам слова, у тому числі формам різних слів, у ході аналізу необхідно робити згортку речення для всіх можливих варіантів граматичних форм. Ті граматичні форми, що забезпечують максимальну згортку дерева (мінімальне число висячих вершин), варто вважати найбільш достовірними.

За результатами практичних досліджень [14], для зняття близько 90% омонімії не потрібно виконувати повний синтаксичний аналіз з подальшою повною згорткою

дерева. Достатнім виявляється використання правил узгодження слів в іменних і дієслівних групах, згортки однорідних членів, узгодження підмета і присудка, прийменниково-відмінкового керування та декількох інших – усього близько 20 правил, описаних безконтекстною граматикою. Докладно ознайомитися зі способами формального опису мови можна, наприклад, у роботі [15].

Побудова інформаційних портретів тексту. Інформаційним портретом тексту називають процес виділення ключових мовних конструкцій. Основною проблемою, що виникає при формуванні інформаційного портрету тексту, є проблема виділення іменних груп – сталих словосполучень, у які входять іменники й погоджені з ними прикметники (наприклад, «розвиток сільського господарства»). Саме цільні іменні групи, а не окремі слова, характеризують зміст тексту і можуть служити для тематичного індексування, авторубрикації і т.д. Супутнім завданням є ранжування значимості іменних груп в тексті – обчислення «тематичної ваги», що вказує на внесок відповідного поняття у зміст тексту (його інформативність).

У ході повного синтаксичного розбору фрази можливе встановлення синтаксичних ролей іменних груп у реченні, що дозволяє ранжувати їх з точки зору важливості для автора фрази. Так, найбільш важливими є слова з групи підмета, потім – присудка, прямого доповнення, непрямого доповнення, обставини. Разом з алгоритмами статистичного аналізу ці факти сприяють більш точному ранжуванню понять за значимістю в інформаційному портреті документа.

Завдання виявлення підмета в більшості випадків вирішуються досить просто й однозначно за рахунок введення відповідних правил узгодження іменника з дієсловом. Завдання виявлення доповнень і обставин вимагає залучення словника моделей керування дієслів, що на даний момент відсутній на ринку в обсязі, достатньому для роботи.

Аналіз тональності тексту. Тональністю тексту будемо називати позитивне або негативне відношення його автора до заданого об’єкта (персони або організації), що фігурує в тексті. Попит на таку технологію з’явився на ринку у зв’язку з активним розвитком політтехнологій та технологій комп’ютерної розвідки. Технологія [16] дозволяє розпізнати:

1. Явну характеристику об’єкта, його дій та їх результатів з використанням тонально-пофарбованої лексики, що несе в собі оціночну семантику, наприклад: X – поганий керівник; геніальний авантюрист X; боязкі дії X; нерішучий X; цинічність дій X; X бездумно погодився; X прийняв авантюрне рішення.

2. Неявну характеристику об’єкта, пов’язану зі згадуванням у тексті таких подій (емоційно-конотативних), пов’язаних з об’єктом, при сприйнятті яких виникає емоційна реакція виду «добре/погано». Наприклад, X бореться з олігархами; дії X привели до росту цін; пенсіонери виступають проти X.

Технологія містить наступні базові операції роботи з текстом:

a) розпізнання всіх згадувань про цільовий об’єкт у тексті, включаючи його повні, короткі, непрямі, займенникові та інші позначення;

b) відсівання та повний синтаксичний розбір тих конструкцій, у яких відбиваються всі ситуації (події і ознаки), пов’язані з цільовим об’єктом;

c) виділення та класифікація тих конструкцій, у яких явно виражається тональність, і тих конструкцій, які описують емоційно-конотативні події.

d) для кожної конструкції ухвалюється рішення про тональність «позитив/негатив» з урахуванням тих місць, які займають у її складі емоційно-конотативні, тональні та нейтральні слова, а також засоби вираження заперечення та інверсії змісту (X нібито не відмовився від авантюрної ідеї).

Семантичний підхід

Проведений нами огляд лінійки продуктів та опрацювання статей даної тематики дозволяє зробити висновок: як правило, під терміном «семантичний аналіз» автори лінгвістичних систем мають на увазі певне перетворення структури вхідного тексту у внутрішню модель представлення даних певної системи. Такі внутрішні моделі зазвичай мають назви «семантична мережа», «семантичний граф», «семантичне представлення» і т.д. Подальша робота типових семантичних алгоритмів базується на порівнянні «семантичної структури» нових документів зі структурою, що отримана системою під час навчання і знаходиться в її базі знань.

Зрозуміло, що даний підхід не можна вважати строго семантичним – націленим на розумну, інтелектуальну обробку тексту, коли автоматизована система насправді розуміє контекст, а не проводить порівняння двох структур, побудованих на базі статистично зібраної інформації про документи. Незважаючи на певну обмеженість згаданих методів, ми проведемо огляд найбільш практично успішних з них.

Первинний семантичний аналіз. Автор методу підкреслює [17], що назва «семантичний» є умовною, так називається метод, в якому використовується валентна структура, описана в словнику РОСС [18]. Семантичний аналіз будує семантичну структуру речень російською мовою. Семантична структура складається із семантичних вузлів і семантичних відносин.

Семантичний вузол – це такий об'єкт текстової семантики, в якого заповнені всі валентності: як експліцитно виражені в тексті, так і імпліцитно – ті, які виходять із екстралінгвістичних джерел.

Семантичне відношення – це універсальний зв'язок, який визначається носієм мови у тексті. Цей зв'язок бінарний, тобто він йде від одного семантичного вузла до іншого.

Семантичні вузли утворюються зі слів вихідного речення. Головне джерело гіпотез про склад семантичного вузла дає синтаксичний аналіз. Більшість синтаксичних груп можуть перейти в семантичні вузли, деякі повинні перетворитися в атрибути вузлів. Крім самого тексту, джерелами гіпотез виступають словник тимчасових груп, словник РОСС [18] та інші тезауруси.

Синтактико-семантичний підхід. В основі підходу лежить лінгвістична модель. Відповідно до цієї моделі основу семантичної структури висловлювання представляє так званий пропозиційний компонент плану змісту. Цей компонент відбиває позамовну ситуацію, що описується реченням, і характеризує його об'єктивний зміст, на відміну від інших компонентів (модального, комунікативного), які так чи інакше характеризують або відношення мовця до ситуації, або співвідносять ситуацію з якимось моментом часу або умовами її реалізації, і тому відносяться до сфери суб'єктивного.

Таким чином, синтактико-семантичний підхід до отримання знань припускає виділення зі структури фрази її семантичного ядра – об'єктивного опису ситуації, і абстрагування від несуттєвих, суб'єктивних компонентів плану змісту. З цією метою використається синтаксичний аналізатор тексту, що працює на підставі знання загальних правил граматики мови, а також словник моделей керування, що описує для кожного предиката способи вираження в мові його аргументів (прийменників та відмінків актантів) [10].

Семантичний аналіз в системі «Мінерва». Виконується у вигляді перетворення графа речення (отриманого після роботи синтаксичного аналізатора) у вирази на внутрішній мові «Мінерва». Функціонує на базі аналізатора та бібліотеки шаблонів синтаксичних конструкцій російської мови, для яких уже створено опис формальною мовою подання знань [19].

Семантика в пошуковій машині. В даному випадку [20] семантичний аналіз тексту має своєю метою витяг змісту з тексту та відображення його у формальну модель, що дозволяє знаходити змістовну близькість двох текстів (для задачі пошуку – близькість запиту та документу). При семантичному аналізі тексту множина синтаксем кожного речення відображається в неоднорідну семантичну мережу, запропоновану Г.С. Осиповим, з синтаксемами у вершинах та семантичними зв'язками як ребрами.

Семантичний аналіз тексту оперує в основному іменними синтаксемами, які виділяються в результаті морфологічного та синтаксичного аналізу. Іменна синтаксема представляється в тексті іменною або прийменниковою групою – словосполученням з іменником або прийменником як керуючим словом. Іменна синтаксема характеризується морфологічною формою – прийменником, відмінком і категоріально-семантичним класом іменника, від якого вона утворена. Морфологічна форма синтаксеми й категоріально-семантичний клас визначаються за допомогою лінгвістичного аналізатора тексту. Синтаксема характеризується також синтаксичною функцією, що вона може виконувати в реченні, і синтаксичним значенням. У ході семантичного аналізу тексту необхідно встановити значення іменних синтаксем, які є носіями змісту тексту.

Морфологічна форма та категоріально-семантичний клас іменної синтаксеми не однозначно задають її значення, тому для вирішення неоднозначностей використовується контекст – дієслово або віддієслівний іменник, при якому іменна синтаксема входить в речення. Для розбору таких випадків використовується спеціальний словник, що описує найбільш часті сполучення певного дієслова з можливими синтаксемами.

Висновки

Дане дослідження проводилося в рамках проекту UWN [21] і було направлено на створення чіткої класифікації лінгвістичних методів та визначення класів задач, що можуть бути вирішені в рамках кожного з методів. За результатами проведеної роботи було визначено чотири основних класи методів, що сьогодні успішно застосовуються на практиці: морфологічний, статистичний, синтаксичний, семантичний. Для кожного з класів було подано перелік типових застосувань та наведено назви прикладних систем, що їх застосовують.

Результати дослідження було використано в проекті UWN при створенні оптимальної структури онтологічної бази знань для використання в семантичних методах. Також, на основі дослідження методів морфологічного підходу в проекті UWN проходить створення системи лексико-граматичного аналізу.

Literatura

1. Yermakov A.Y. Kompyuternaya lingvistika i intellektual'nye tehnologii. Vyp. 8 (15). M: RGGU. 2009. 620 s.
2. Francis W. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Brown University Providence, Rhode Island Department of Linguistics Brown University. 1979
3. Mihailyan A. Nekotorye metody avtomaticheskogo analiza yestestvennogo yazyka, ispol'zuemye v promyshlennyh produktah. http://www.citforum.ru/programming/digest/avtestlang.shtml
4. Van Guilder L. Automated Part of Speech Tagging: A Brief Overview (Handout for LING361, Fall 1995 Georgetown University). Georgetown University. 1995
5. Brill E. Proceedings of ACL-95. 1995
6. Tapanainen P. Computational and Language E-print Archive. 1994
7. Samuelsson Ch. Proceedings of 35 Annual Meeting of the Association for Computational Linguistics and 8th conference of the European Chapter of the Association for Computational Linguistics. ACL. Madrid. 1997

8. Volk M. Proceedings of KONVENS-98. Bonn. 1998.
9. Yermakov A.Y. Informatsionnye tehnologii. №12. 2000
10. Yermakov A.Y. Tezisy doklada mezhdunarodnogo kongressa "Russkii yazyk: istoricheskie sud'by i sovremennost'"
11. Klimenkov S.V. Nauchno-tehnicheskii vestnik, vypusk 46: informatsionnye i telekommunikatsionnye sistemy. Sankt-Peterburgskii Gosudarstvennyi universitet informatsionnyh tehnologii, mehaniki i optiki. 2008. S 198-202
12. Kupiec J. A Trainable Document Summarizer - Xerox Palo Alto Research Centre, Palo Alto, CA. 1995
13. Harlamov A.A. Informatsionnye tehnologii. № 2. 1998
14. Yermakov A.Ye. Informatsionnye tehnologii. № 7. 2002
15. Gladkii A.V. Formal'nye grammatiki i yazyki. M.: Nauka. 1973
16. Yermakov A.Ye. Komp'yuternaya lingvistika i intellektual'nye tehnologii: trudy Mezhdunarodnoi konferentsii Dialog'2005. Moskva: Nauka. 2005
17. Sokirko A. V. Semanticheskie slovari v avtomaticheskoi obrabotke teksta: Po materialam sistemy DIALING. M. 2001. 120 s.
18. Lyeont'eva N.N. Russkii obshchesemanticheskii slovar' (ROSS): struktura, napolnenie. NTI. Ser. 2. 1997. № 12. S 5-20
19. Proekt "Minerva" http://www.inteltec.ru/publish/articles/textan/concept.shtml
20. Tihomirov I.A Materialy mezhdunarodnoi konferentsii "Dialog 2009"
21. UWN project. http://lingvoworks.org.ua/

А.А. Никоненко

Обзор компьютерно-лингвистических методов обработки естественноязыковых текстов

В статье проведено исследование методов автоматического анализа естественноязыковых текстов. Подробно рассмотрены четыре основных подхода: морфологический, статистический, синтаксический, семантический. Описаны возможности их использования в решении прикладных задач и специфика применения. В качестве примеров приведен перечень прикладных систем, использующих перечисленные методы.

A.A. Nykonenko

Overview of Computer-Linguistic Methods for Processing of Natural Language Texts

The paper studied the methods for automatic analysis of natural language texts. Four main approaches are discussed in detail, i.e. morphological, statistical, syntactic, semantic. The possibility of their use in solving practical problems and specificity of application are discussed. As the examples the list of application systems, which use these methods, are given.

Стаття надійшла до редакції 31.05.2011.